XXXVIII Jornadas de Automática
https://doi.org/10.17979/spudc.9788497497749.0976

COMPARACIÓN DE MÉTODOS DE DETECCIÓN DE ROSTROS EN IMÁGENES DIGITALES

Natalia García del Prado[1], Víctor González-Castro[1,2], Enrique Alegre[1,2], Eduardo Fidalgo Fernández[1,2]
[1] Departamento de Ingeniería Eléctrica y de Sistemas y Automática, Universidad de León, España
[2] Investigador en INCIBE (Instituto Nacional de Ciberseguridad)
ngarcd00@estudiantes.unileon.es, {vgonc, ealeg, efidf }@unileon.es

Resumen

En este trabajo se realiza la evaluación de tres métodos de detección de rostros con cuatro conjuntos de imágenes utilizados en la literatura relacionada con este problema. Los métodos evaluados son el método de detección de objetos de Viola & Jones, un método basado en una modificación de HOG implementado en la librería DLib, y un método basado en Redes Neuronales Convolucionales llamado Multi-task Cascaded Convolutional Neural Networks (MTCNN). Los resultados obtenidos con los conjuntos de datos evaluados indican que el método que mejores resultados globales ha obtenido ha sido MTCNN, tanto en términos de FScore como en recall, i.e. tasa de verdaderos positivos.

Palabras clave: Detección de rostros, Viola & Jones, DLib, MTCNN.

1. INTRODUCCIÓN

Durante los últimos años se ha producido un gran auge en aplicaciones de detección de rostros utilizando procesamiento de imágenes digitales y reconocimiento de patrones. Una de las razones que ha llevado a este crecimiento son las necesidades cada vez mayores de esta tecnología en aplicaciones de seguridad y vigilancia utilizadas en diferentes ámbitos. Pero también en el campo de las aplicaciones de dispositivos móviles está siendo de vital importancia ya que muchas de estas apps utilizan la detección de rostros en una fotografía o en un video (p. ej., Snapchat).

Existe un gran número de áreas en las que se desarrollan programas para la detección automática de rostros, bien con el único objetivo de la detección o como paso previo al reconocimiento de los mismos, y requieren la mayor precisión posible. En el campo de la seguridad, por ejemplo, la detección automática de rostros puede ser de gran ayuda para los investigadores ya que permite el filtrado automático de imágenes o videos en función de la presencia o no de rostros.

El principal objetivo de este artículo es el estudio de varios métodos del estado del arte en detección de rostros y la evaluación de los resultados obtenidos al ejecutarlos contra varios conjuntos de imágenes para verificar cuál de ellos es el que mejor funciona y el que detecta mayor número de rostros. Estos resultados pueden aportar mucha información previa al desarrollo de programas que impliquen una detección precisa del rostro, pues permitirá implementar el método más adecuado.

Existen muchos campos en los que se aplica la detección facial, tales como la extracción de contenido de imagen, codificación de video, videoconferencia, etc., aplicaciones de videovigilancia e interfaces computadora-humano. Otro campo a destacar donde se utiliza la detección de rostros son los sistemas de autenticación biométrica, como los sistemas de autenticación mediante reconocimiento facial, que se investigan activamente para las aplicaciones de control de acceso y seguridad. Además, la detección de rostros también se emplea en el área del entretenimiento (p. ej., videojuegos, realidad virtual, álbumes de fotos, maquillaje virtual, aplicaciones de Smartphone, etc.).

En este trabajo se han evaluado diferentes métodos de detección de rostros: (a) el algoritmo de Viola-Jones [1], (b) el método de detección de rostros propuesto por Kazemi y Sullivan [2] implementado en la librería Dlib y (c) el método Multi-task Cascaded Convolutional Neural Networks (MTCNN) [3]. En esta evaluación se han utilizado varios conjuntos de datos publicos: FDDB [4], WIDER FACE [5], MALF [6] y AFW [7].

El resto del artículo se estructura de la siguiente manera: En la Sección 2 se realiza un breve repaso a los antecedentes y el estado del arte de métodos de detección de rostros. A continuación, en la Sección 3 se explican brevemente los métodos que se han evaluado y los conjuntos de datos usados en dicha evaluación. Los experimentos y sus resultados son explicados y discutidos en la Sección 4 y, finalmente, las conclusiones se exponen en la Sección 5.

2. ESTADO DEL ARTE

La detección de rostros se lleva desarrollando desde hace varias décadas. En la de 1950 se realizaron los primeros experimentos sobre detección de rostros desde el punto de vista de la psicología [8]. Tras estos se realizaron otras investigaciones como las interpretación de las diversas expresiones de la cara, la interpretación de las emociones o la percepción de gestos.

Durante las décadas de los 70 y 80 del siglo pasado se utilizaban plantillas y mediciones de características geométricas de partes del rostro para detectar y reconocer caras [9]. El estudio sobre la detección de rostros continuó desde entonces: por ejemplo, en 1994 Yang y Huang [10] propusieron un método jerárquico basado en conocimiento que va refinando la detección del rostro en cada etapa. Li y colaboradores presentaron un método para detectar y seguir rostros en vídeos en color utilizando detección de la piel a través del color junto con un modelo facial para detectar rostros dentro de regiones de la piel [11]. Viola y Jones presentaron un método de detección de objetos [1] que también serviría para detección de rostros que, aún hoy, es muy utilizado. Por ejemplo, para aplicaciones como videovigilancia e interfaces computadora-humano, Kim y colaboradores propusieron en el año 2002 un método para detectar y rastrear caras [12]. Este método se aplicó en la región candidata extraída por el movimiento, el color y la información de aspecto global. A continuación, las características extraídas de estas regiones candidatas mediante análisis de componentes independientes (ICA en sus siglas en inglés) se clasificaron mediante máquinas de vectores de soporte (SVM en sus siglas en inglés). Los experimentos se realizaron tanto en imágenes en escala de gris como en secuencias de vídeo en color, obteniendo tasas de detección del 91 %, con muy pocas falsas alarmas, i.e. falsos positivos.

Con la aparición del Aprendizaje Profundo (*Deep Learning*) se han comenzado a utilizar Redes Neuronales Convolucionales a la detección de rostros, con muy buenos resultados [3, 13, 14].

3. MÉTODOS

En esta sección se describirán brevemente los métodos evaluados en el presente trabajo.

3.1. ALGORITMO VIOLA & JONES

El algoritmo clásico de Viola & Jones [1] es uno de los algoritmos más utilizado en la detección de rostros debido a su robustez y velocidad. Trabaja con funciones que, mediante el uso de regiones rectangulares de diversos tamaños, buscan el rectángulo más pequeño que es el que contendrá la cara que se busca.

Se basa en extraer características de la imagen basadas en la sustracción de los valores de intensidad de los pixels de la imagen en una región rectangular. Específicamente se resta la suma de los niveles de gris de los pixels que se encuentran bajo la parte blanca del rectángulo menos los que se encuentran bajo la parte negra (ver ejemplos en la Figura 1). Estas características se llaman características Haar pues los rectángulos empleados son una reminiscencia de las funciones base de Haar [1].

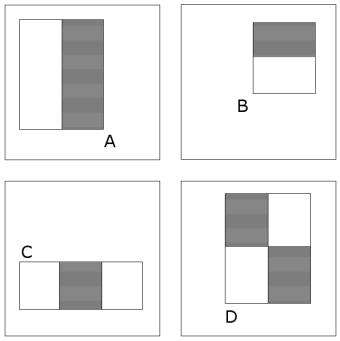

Figura 1: Ejemplos de regiones rectangulares empleadas en la extracción de las características Haar.

El cálculo de estas características se puede realizar rápidamente gracias a la utilización de la *imagen integral* [1]. Dada una imagen i, su integral I en el punto (x, y) se calcula sumando todos los pixels a la izquierda y por encima de (x, y) (ver Ecuación (1)).

$$I(x,y) = \sum_{x' \leq x, y' \leq y} i(x', y') \qquad (1)$$

El número de características así calculadas incluso para una región pequeña es muy alta, y solo un pequeño conjunto de las mismas es útil para la detección. Por ello, se emplea Adaboost para reducir el número de características.

Por último, se utiliza una cascada de clasificadores para evaluar las diferentes ventanas, de modo que si alguno de los clasificadores de la cascada falla, se descarta la existencia de un rostro en la subventana evaluada, de manera que la velocidad del método se incrementa.

3.2. ALGORITMO DE DETECCIÓN DE ROSTROS IMPLEMENTADO EN LIBRERÍA DLIB

Este detector está basado en un clasificador que utiliza características basadas en una variante [15]

de histogramas de gradientes orientados (HOG) [16] extraidas de ventanas deslizantes de tamaño fijo operando sobre pirámides de imágenes.

Una vez detectado el rostro se puede utilizar el método de Kazemi y Sullivan [2] para detectar la pose de la cara mediante la detección de puntos característicos de la misma, o *landmarks* (ver Figura 2). Sin embargo, en este trabajo solo se ha utilizado para detectar caras en posición frontal para compararlo con el método de Viola & Jones.

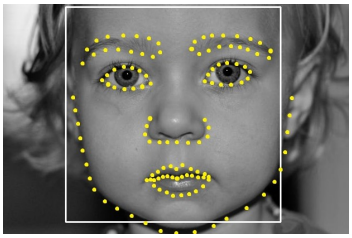

Figura 2: Ejemplos de puntos característicos detectados en un rostro.

3.3. MULTI-TASK CASCADED CONVOLUTIONAL NEURAL NETWORKS (MTCNN)

Zhang y colaboradores propusieron la utilización de una cascada de Redes Neuronales Convolucionales para detectar rostros y *landmarks* en los mismos [3], que consta de tres etapas. Antes de comenzar, se redimensiona la imagen a diferentes escalas para construir una pirámide de imágenes. En la primera etapa se utiliza una red convolucional que detecta ventanas de caras candidatas. A continuación, se utiliza otra red neuronal convolucional que descarta un gran número de candidatos en los que no existen rostros. Finalmente una última red convolucional trata de identificar en cuáles de los candidatos existe realmente un rostro, identificando las posiciones de cinco puntos de referencia faciales: uno en cada ojo, otro en la punta de la nariz y los dos restantes en las comisuras de los labios.

4. EXPERIMENTOS

4.1. DATASETS UTILIZADOS

4.1.1. FDDB: Face Detection Data Set Benchmark

FDDB es un conjunto de imágenes de regiones faciales diseñada para estudio de los problemas de la detección de rostros sin restricciones [4]. Este conjunto de datos contiene las anotaciones para 5171 caras en un subconjunto del dataset "Faces in the Wild" que consta de 2845 imágenes.

4.1.2. WIDER FACE

El conjunto de datos WIDER FACE [5] consta de 32.203 imágenes con 393.703 caras anotadas con un alto grado de variabilidad en escala, pose y oclusión extraídas del conjunto de datos público WIDER. El conjunto de imágenes WIDER FACE se organiza en 61 clases de eventos muy diversos, como imágenes de desfiles de bandas, manifestaciones, conferencias de prensa, reuniones familiares o incluso funerales. En este trabajo se ha utilizado un subconjunto de imágenes que contienen unas 40200 caras anotadas.

4.1.3. MALF: Multi-Attribute Labelled Faces

MALF es un conjunto abreviado de imágenes cuyas etiquetas contienen múltiples atributos faciales recopilados y proporcionados por el Centro de Biometría y de Investigación de Seguridad [6]. MALF es el primer conjunto de datos de detección de rostros que soporta una evaluación fina ya que contiene 5.250 imágenes tomadas en entornos reales con un total de 11.931 caras. Sin embargo, en este trabajo se ha utilizado un subconjunto de 250 imagenes para las que se tenía la anotación del *ground truth* que contienen unos 600 rostros.

4.1.4. AFW: Annotated Facial in the Wild

AFW [7] contiene 205 imágenes con 473 caras etiquetadas. Para cada cara, las anotaciones incluyen un rectángulo circunscrito (*bounding box*), 6 puntos de referencia y los ángulos de la postura. Este dataset, por lo tanto, permite evaluar métodos tanto de detección de rostros como de estimación de posición y de hitos en imágenes desordenadas del mundo real.

4.2. EXPERIMENTOS Y RESULTADOS

Se han probado los tres métodos indicados en las secciones 3.1, 3.2 y 3.3 con cada uno de los cuatro conjuntos de imágenes comentados en la sección 4.1. Cuando el método evaluado detecta un rostro, el *bounding box* que devuelve se compara con la de la *ground truth* del conjunto de imágenes. La comparación se realiza evaluando la proporción de la superficie de interesección de ambos bounding boxes sobre el área total del ground truth, de modo que si esta es mayor que un determinado umbral, se considera el rostro como bien detectado. En caso contrario se considera que el rostro no se detecta. El umbral que se ha establecido en este trabajo ha sido 0.7.

Sean TP, FP y FN el número de verdaderos po-

sitivos (i.e., el número de rostros detectados correctamente), falsos positivos (i.e., el número de veces que se detecta un rostro donde en realidad no existe) y falsos negativos (i.e., el número de rostros que no se detectan), respectivamente. Las métricas de error utilizadas en este trabajo han sido:

- *Precision*: Indica la tasa de rostros que detecta correctamente entre todos los elementos que detecta como un rostro.

$$precision = \frac{TP}{TP + FP} \tag{2}$$

- *Recall*: Indica la tasa de rostros detectados correctamente detectados sobre el total de rostros que se encuentran realmente en las imagenes.

$$recall = \frac{TP}{TP + FN} \tag{3}$$

- *FScore*: Sean P y R la *precision* y *recall*, respectivamente. El FScore es la media harmónica de P y R.

$$FScore = 2\frac{P \cdot R}{P + R} \tag{4}$$

A continuación se discuten los resultados obtenidos con cada uno de los datasets evaluados.

4.2.1. Resultados con el dataset FDDB

Se han evaluado los tres métodos con todas las imágenes del conjunto FDDB (i.e., 2845 imágenes que contienen 5171 rostros). En la Tabla 1 se muestran los resultados utilizando un umbral para la detección del rostro de 0.7.

Tabla 1: Resultados de los métodos evaluados con el dataset FDDB.

	Recall	**Precision**	**FScore**
Viola & Jones	0.63	0.92	0.75
DLib	0.56	**0.99**	0.71
MTCNN	**0.77**	0.93	**0.84**

Globalmente, el método MTCNN es el mejor de los tres evaluados con este dataset ya que el FScore que consigue mejora al del obtenido por los otros dos métodos. La tasa de verdaderos positivos (i.e. *recall*) es notablemente superior respecto a los otros métodos, lo que indica que es el que más rostros detecta correctamente. Además, el hecho de que obtenga menor *precision* que el método incluido en la librería DLib indica que detecta más elementos de la imagen como rostros que no son rostros (i.e., más falsos positivos) que este.

En la Figura 3 se muestran los *recalls* de los diferentes métodos en función de los diferentes umbrales evaluados (i.e. 0.1, 0.3, 0.5, 0.7 y 0.9).

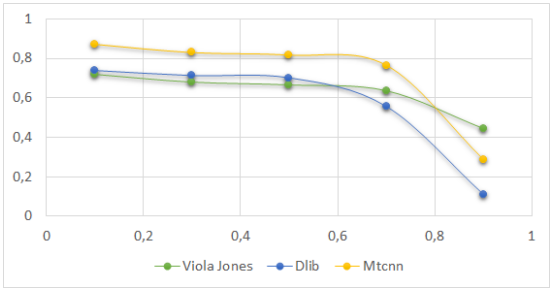

Figura 3: *Recall* obtenido por los diferentes métodos en el dataset FDDB con umbrales de detección diferentes.

En la Figura 3 se puede comprobar que MTCNN tiene mayor *recall* (i.e., tasa de verdaderos positivos) en todos los escenarios excepto en el más restrictivo (i.e. aquél en el que se exige que el 90 % de la superficie del rectángulo de la cara detectada esté superpuesto al del *ground truth* del *dataset*).

4.2.2. Resultados con el dataset WIDER FACE

También se han evaluado los tres métodos con el conjunto de imágenes WIDER FACE. Como se indicó en la sección 4.1.2 se ha utilzado un subconjunto de imágenes que consta de unas 40200 rostros anotados. En la Tabla 2 se muestran los resultados utilizando un umbral para la detección del rostro de 0.7.

Tabla 2: Resultados de los métodos evaluados con el subconjunto del dataset WIDER FACE.

	Recall	**Precision**	**FScore**
Viola & Jones	0.14	0.86	0.23
DLib	0.11	**0.99**	0.20
MTCNN	**0.36**	0.96	**0.53**

Una vez más, el método MTCNN ha obtenido mejores resultados globalmente que los otros dos métodos (*FScore* de 0.53 y *recall* de 0.36), si bien los resultados con este dataset no han resultado tan buenos como con FDDB.

En la Figura 4 se muestran los *recalls* de los diferentes métodos en función de los diferentes umbrales evaluados (i.e. 0.1, 0.3, 0.5, 0.7 y 0.9).

En el caso de WIDER, MTCNN obtiene un *recall* superior a los otros dos métodos con todos los umbrales de detección evaluados.

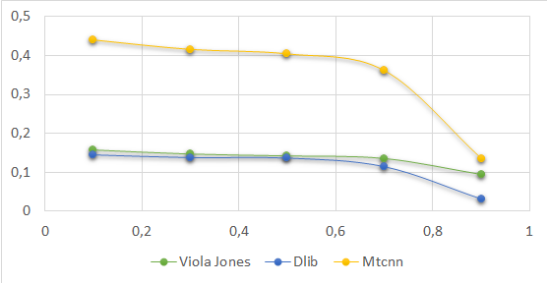

Figura 4: *Recall* obtenido por los diferentes métodos en el dataset WIDER con umbrales de detección diferentes.

4.2.3. Resultados con el dataset MALF

Se han evaluado los tres métodos con un subconjunto de 250 imágenes extradas del conjunto de imágenes MALF. En la Tabla 3 se muestran los resultados utilizando un umbral para la detección del rostro de 0.7.

Tabla 3: Resultados de los métodos evaluados con el dataset MALF.

	Recall	**Precision**	**FScore**
Viola & Jones	0.47	0.86	0.61
DLib	0.48	**0.99**	0.64
MTCNN	**0.57**	0.90	**0.70**

Siguiendo la misma tendencia, el método MTCNN ha obtenido mejores resultados que los otros dos métodos (*FScore* de 0.70 y *recall* de 0.57). En la Figura 5 se muestran los *recalls* de los diferentes métodos en función de los diferentes umbrales evaluados (i.e. 0.1, 0.3, 0.5, 0.7 y 0.9).

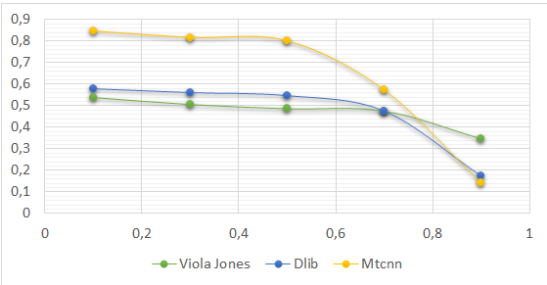

Figura 5: *Recall* obtenido por los diferentes métodos en el dataset MALF con umbrales de detección diferentes.

Una vez más, MTCNN obtiene una tasa de rostros detectados correctamente superior a los otros dos métodos en todos los escenarios excepto en el más restrictivo.

4.2.4. Resultados con el dataset AFW

Finalmente, se han evaluado los tres métodos con con todas las imágenes del conjunto AFW (i.e., 205 imágenes con 473 rostros etiquetados). En la Tabla 4 se muestran los resultados utilizando un umbral para la detección del rostro de 0.7.

Tabla 4: Resultados de los métodos evaluados con el dataset AFW.

	Recall	**Precision**	**FScore**
Viola & Jones	0.67	0.57	0.62
DLib	0.66	**0.98**	0.79
MTCNN	**0.94**	0.69	**0.80**

El comportamiento de los diferentes métodos evaluados es el mismo que en el caso de los otros datasets: MTCNN obtiene mejores resultados globales que los otros, aunque en este caso el FScore de MTCNN y del método implementado en la librerá DLib es similar. En la Figura 6 se muestran los *recalls* de los diferentes métodos en función de los diferentes umbrales evaluados (i.e. 0.1, 0.3, 0.5, 0.7 y 0.9).

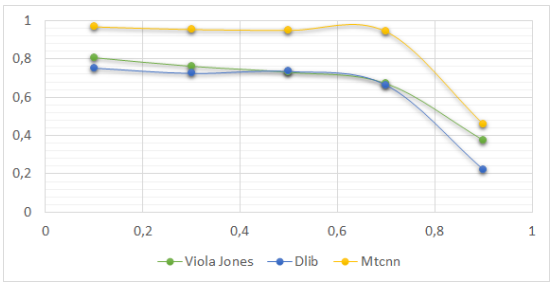

Figura 6: *Recall* obtenido por los diferentes métodos en el dataset AFW con umbrales de detección diferentes.

Siguiendo la misma tendencia, MTCNN obtiene *recalls* superiores a los otros dos métodos para todos los umbrales de detección.

5. CONCLUSIÓN

En este trabajo se han evaluado tres métodos de detección de rostros: el algoritmo de detección de objetos propuesto por Viola & Jones, un método basado en una variante de HOG implementado en la librería DLib y, por último, el método Multi-task Cascaded Convolutional Neural Networks (MTCNN). El rendimiento de estos tres métodos en la tarea de detección de rostros se ha comparado utilizando cuatro conjuntos de imágenes disponibles públicamente diferentes ampliamente utilizadas en la literatura: FDDB, WIDER FACE, MALF y AWF.

En todos los casos el método que mejor rendimiento global ha obtenido ha sido MTCNN. Tanto el FScore como el *recall* (i.e. la tasa de verdaderos positivos) han sido superiores usando MTCNN en todos los conjuntos de datos. Por otro lado, llama la atención que el método implementado en la librería DLib obiene mejor *precision* en todos los casos. Esto se debe a que el número de falsos positivos (i.e. detecciones de caras donde no existen rostros) que detecta este método es menor que el de MTCNN. Por otro lado, el método de Viola & Jones y el implementado en la librería DLib presentan resultados similares en cuanto a *recall* aunque el FScore de este último es algo superior. Una de las mayores limitaciones del método de Viola & Jones es que en este trabajo se ha utilizado solo la casacada de clasificadores Haar para detectar rostros frontales. Es posible que, combinando esta con la cascada para detectar rostros de perfil se lograsen mejorar los resultados de este método.

En futuros trabajos se tratará de aplicar la detección de rostros a diferentes aplicaciones (p. ej., conteo de personas en eventos o transporte público), o aplicar reconocimiento de los rostros detectados o incluso evaluar de manera automática las emociones de una persona determinada mediante el análisis de sus gestos.

Agradecimientos

Esta investigación ha sido financiada por el acuerdo marco entre la Universidad de León e INCIBE (Instituto Nacional de Ciberseguridad) bajo la adenda 22.

Referencias

[1] Paul Viola and Michael Jones. Rapid object detection using a boosted cascade of simple features. In *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, volume 1, pages I–I. IEEE, 2001.

[2] Vahid Kazemi and Josephine Sullivan. One millisecond face alignment with an ensemble of regression trees. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1867–1874, 2014.

[3] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23 (10):1499–1503, 2016.

[4] Vidit Jain and Erik G Learned-Miller. Fddb: A benchmark for face detection in unconstrained settings. *UMass Amherst Technical Report*, 2010.

[5] Shuo Yang, Ping Luo, Chen-Change Loy, and Xiaoou Tang. Wider face: A face detection benchmark. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5525–5533, 2016.

[6] Bin Yang, Junjie Yan, Zhen Lei, and Stan Z Li. Fine-grained evaluation on face detection in the wild. In *Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on*, volume 1, pages 1–7. IEEE, 2015.

[7] Xiangxin Zhu and Deva Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2879–2886. IEEE, 2012.

[8] Jerome S Bruner and Renato Tagiuri. The perception of people. Technical report, DTIC Document, 1954.

[9] Mark Nixon. Eye spacing measurement for facial recognition. In *29th Annual Technical Symposium*, pages 279–285. International Society for Optics and Photonics, 1985.

[10] Guangzheng Yang and Thomas S Huang. Human face detection in a complex background. *Pattern Recognition*, 27(1):53 – 63, 1994. ISSN 0031-3203.

[11] Y. Li, A. Goshtasby, and O. Garcia. Detecting and tracking human faces in videos. In *Proceedings 15th International Conference on Pattern Recognition. ICPR-2000*, volume 1, pages 807–810 vol.1, 2000.

[12] Tae-Kyun Kim, Sung-Uk Lee, Jong-Ha Lee, Seok-Cheol Kee, and Sang-Ryong Kim. Integrated approach of multiple face detection for video surveillance. In *Pattern Recognition, 2002. Proceedings. 16th International Conference on*, volume 2, pages 394–397. IEEE, 2002.

[13] Qin-Qin Tao, Shu Zhan, Xiao-Hong Li, and Toru Kurihara. Robust face detection using local cnn and svm based on kernel combination. *Neurocomputing*, 211:98 – 105, 2016. ISSN 0925-2312. SI: Recent Advances in SVM.

[14] Shaohui Lin, Ling Cai, Xianming Lin, and Rongrong Ji. Masked face detection via a modified lenet. *Neurocomputing*, 218:197 – 202, 2016. ISSN 0925-2312.

[15] Pedro F Felzenszwalb, Ross B Girshick, David McAllester, and Deva Ramanan. Object detection with discriminatively trained part-based models. *IEEE transactions on pattern analysis and machine intelligence*, 32(9): 1627–1645, 2010.

[16] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 886–893. IEEE, 2005.